DOCUMENT RESUME

ED 021 479 56 EM 006 296

By- Rosen, Ellen F.; Stolurow, Lawrence M.

A METHOD FOR ESTIMATION OF DIFFICULTY AND DISCRIMINATION INDICES IN PROGRAMED LEARNING. COMPARATIVE STUDIES OF PRINCIPLES FOR PROGRAMMING MATHEMATICS IN AUTOMATED INSTRUCTION. TECHNICAL REPORT NUMBER 10.

Illinois Univ., Urbana

Spons Agency- Office of Education (DHEW), Washington, D.C.

Report No- NDEA-7A-806

Pub Date Jul 64

Contract- OEC-711151-01

Note- 12p.

EDRS Price MF-$0.25 HC-$0.56

Descriptors- *DATA ANALYSIS, *ITEM ANALYSIS, *MATHEMATICAL APPLICATIONS, *PROGRAMED INSTRUCTION, PROGRAMED MATERIALS, STATISTICS, *TEST CONSTRUCTION

In the development of programed materials, usually all responses are analyzed for a small sample of students. The viable alternative is to perform mathematical transformations on response data from a selected portion of a large sample (more than 400 students) to obtain reliable estimates of item difficulty and item discrimination. (LH)

UNIVERSITY OF ILLINOIS
Urbana, Illinois

A Method for Estimation of Difficulty and Discrimination Indices in Programed Learning

Lawrence M. Stolurow and Ellen F. Rosen

COMPARATIVE STUDIES OF PRINCIPLES FOR PROGRAMMING MATHEMATICS IN AUTOMATED INSTRUCTION

Technical Report No. 10
July, 1964

Co-Investigators:

Lawrence M. Stolurow
Professor, Department of Psychology
Training Research Laboratory

Max Beberman
Professor, College of Education
University of Illinois Committee
on School Mathematics (UICSM)

Project Sponsor:

Educational Media Branch
U. S. Office of Education
Title VII

Project No. 711151.01

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

U. S. Office of Education
Title VII

COMPARATIVE STUDIES OF PRINCIPLES FOR PROGRAMMING MATHEMATICS IN AUTOMATED INSTRUCTION

ED021479

EM 006 296

Technical Report No. 10.

A Method for Estimation of Difficulty and Discrimination Indices in Programed Learning.

Ellen F. Rosen and Lawrence M. Stolurow

July, 1964

A Method for Estimation of Difficulty and Discrimination Indices in Programed Learning

Ellen F. Rosen and Lawrence M. Stolurow
University of Illinois

When using large samples of subjects in a test development program, one is faced with the problem of data analysis. Often it is not economically feasible to actually use all of the data which has been gathered. In view of this, an accurate method of estimating the value of the desired statistic for the whole sample from a smaller portion of the sample is desirable.

In test analyses, the statistics usually used for revision purposes are the difficulty index and the discrimination index (Wood, 1961). The former is the proportion of subjects who answer the item correctly. Thus, a high difficulty index indicates less difficult items. The discrimination index is a measure of how well the test item discriminates subjects with respect to some criterion, i.e., how well the item correlates with some criterion measure. In essence, the discrimination index is a measure of the validity of the item; a measure of how well it predicts the criterion.

PROBLEM

In test construction, usually an item difficulty of 50% is aimed for. In programed learning it is possible to consider each frame (each question) as a test item, the difference being that the feedback (knowledge of results) given to the student is relatively immediate, and that the purpose of the item is not the same as a similar test item. In programed learning, the situation

is constructed in order to foster learning; in a test, the situation is constructed in order to determine how much learning has already occurred. Thus, the question of what the optimum level of difficulty is becomes less settled. Should most students be unable to answer the item or should most students be allowed to succeed? Item difficulty level is a variable in the field of programed learning; the optimum level of difficulty is yet to be empirically determined. Further, item difficulty is intimately related to the concept of step size. A possible empirical measure of the step size from item (a) to item (b) is the difference in difficulty of the two items. Thus, the estimate of total sample item difficulty should be scaled so that it has the property of additivity.[1]

Discrimination Index

Kelley (1939) attacked the problem of finding the best sample for estimating the discrimination of a test item for the whole sample. He made the following assumptions:

(1) that if 2j individuals are to be selected as the sample on which the estimate is to be based, the best results will be achieved if j individuals are chosen at the bottom of the distribution and j at the top;

(2) that the whole sample is of size N, where N = 2m;

(3) that the scores are graduated;

(4) that the scores are normally distributed;

---

[1] For additional discussion of issues relating to both tests and problems see Jacobs (1962).

(5) that the certainty with which these two groups are differentiated is given by

(i) $$f(j) = \frac{\bar{x}_u - \bar{x}_l}{S_{\bar{x}_u - \bar{x}_l}}$$, where $\bar{x}_u$ is the mean deviation score of the upper group and $\bar{x}_l$ is the mean deviation score of the lower group.[2]

A pictorial representation of the situation is given in Figure 1. The problem has now been reduced to the mathematical problem of solving equation (i) for j such that f is maximized. However, before differentiating, the scores need to be corrected for systematic error in order to work with the "true scores." This systematic error arises as a consequent of the particular sampling method employed -- a score that is not randomly selected, but rather chosen because it has a certain deviation will suffer from a systematic error. This error can be corrected for by regressing the score toward the mean:

$x'_a = rx_a$ where r is the reliability coefficient.

The standard deviation of x'_a is $S\sqrt{r-r^2}$ where S is the standard deviation of the 2m measures. All the other predicted scores will have the same standard deviation.

Thus $\bar{x}'_j = \sum_1^j x'_j = \sum_1^j \cdot rx_j = r\bar{x}_j$, and $S_{\bar{x}'_j} = \frac{S_j}{\sqrt{j}} = \frac{S\sqrt{r-r^2}}{\sqrt{j}}$

Thus, the critical ratio f(j) to be maximized becomes:

$$f(j) = \frac{\bar{x}'_u - \bar{x}'_l}{S_{\bar{x}'_u - \bar{x}'_l}} = \frac{2r\bar{x}'_j}{S_{2r\bar{x}_j}} = \frac{\sqrt{2r\bar{x}_j}\sqrt{j}}{S\sqrt{r-r^2}}, \quad r > 0.$$

---

[2] f is the statistic commonly known as the "critical ratio." It is clear that as the two groups become more clearly discriminable the distance between $\bar{x}_u$ and $\bar{x}_l$ will get larger, but $\sigma_{\bar{x}_u - \bar{x}_l}$ will remain unchanged.

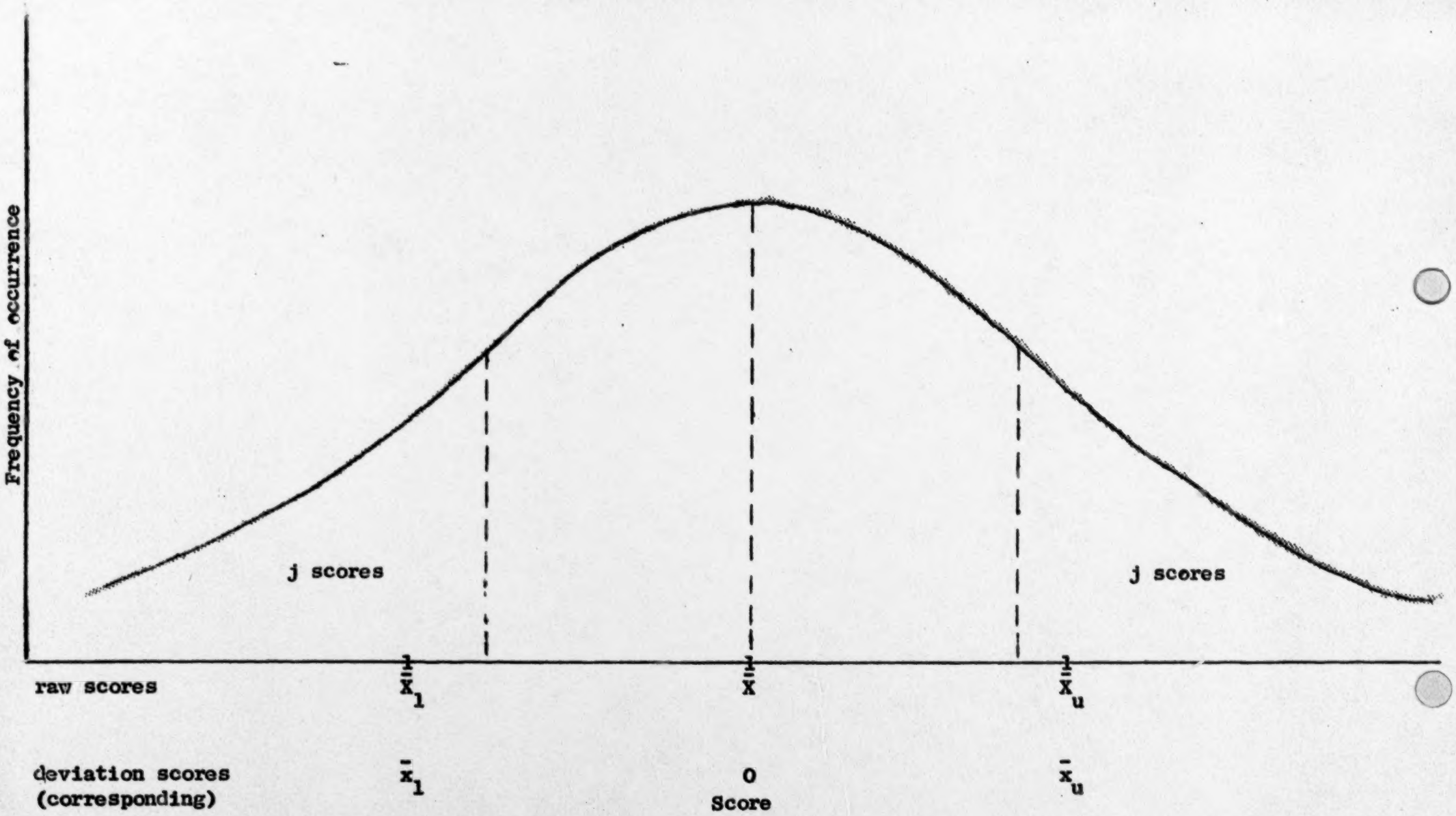

Fig. 1. Hypothetical distribution of N scores. The scores increase from left to right.

Since

$$f(j) = K\sqrt{j}\ \bar{x}_j, \text{ where } K = \frac{\sqrt{2r}}{s\sqrt{r-r^2}},\ r > 0,$$

then

$f(j)$ is a maximum when $\sqrt{j}\ \bar{x}_j$ is a maximum.

Let $q = j/N$, then $\bar{x}_j = z/q$ where z is the ordinate of the standard normal distribution (0, 1) at the point x where x is the end point of the tail which contains the proportion q of the cases of the distribtuion. Thus

$$f(x) = K\sqrt{qN}\ \left(\frac{z}{q}\right) = K\sqrt{N}\ z/q = \frac{Cz}{\sqrt{q}}$$

and

$$\frac{df}{dx} = C\left(\frac{1}{\sqrt{q}}\ \frac{dz}{dx} - \frac{1}{2}\ \frac{2}{q^{3/2}}\ \frac{dq}{dx}\right.$$

$$= \frac{Cz}{\sqrt{q}}\ (-x + z/2q),\ 0 < q \leq .50.$$

Thus, $-x + z/q = 0$. That is $q = z/2x$. Kelley (1939) asserts that this value makes f a maximum. He further asserts that $q = z/2x$ when $q = .2702678$.

Davis (1949) presents a table for quickly calculating discrimination indices based on the upper and lower 27%. He improves Kelley's work by using Fisher's z as a direct measure of the discriminating power of an item. The advantage in using Fisher's z is that a given increase in the value of Fisher's z has essentially a constant meaning at any part in the range of its possible values. Davis also states that the correlation coefficients obtained by this procedure are not greatly affected by the difficulty levels of the

test items. Thus, if similar samples of students are used to determine the discrimination power of a set of test items, Fisher's z transformation of these correlation coefficients can be legitimately added, subtracted, or averaged.

Davis (1949) reports that the reliability of the discrimination indices calculated using his table (which is entered by means of the proportion of successes on the item in the upper and lower 27% of the sample) based on 100 cases in each tail is approximately 0.60.

Difficulty Index

Davis (1949) also presents a method for determining the difficulty of an item (based on the proportion of success on the item of the upper and lower 27%) which he states leads to a reliability of about .98 for a set of difficulty indices when the size of each tail is about 100 cases. His development is briefly outlined in the next few paragraphs with an adaptation to programed items.

Let P represent the proportion of students of the total sample that know the answer to an item, then P is defined as follows:

$$P = \frac{R - \frac{W}{K - 1}}{N - NR})$$

where R = number of students giving correct answers,

W = number of students giving incorrect answers,

K = number of choices of answers for the item,

N = number of students in the sample,

NR = number of students who do not reach the item in the time limit.

In programed instruction, NR = O, since, with no time limit, all students will have an opportunity to read all questions. Therefore, all ommissions must be considered as errors. It is possible that a student will omit a page he is not supposed to omit, but this must be counted as an error (not following instructions). The expression reduces to

$$P = \frac{R - W/(K-1)}{N}.$$

Furthermore, programed instructional items often will be of the fill-in-the-blank type. The size of the class of responses available to the student must be then determined from the context of the question.

Let u be the set of all existing responses, S be the subset of u consisting of all responses potentially available to the student to use in answering the item, A be the subset of S consisting of the responses to the item, A be the subset of S consisting of the incorrect responses, and m (X) be the size of the set X or the number of elements or responses in the set X, then $K = m(A \cup A') = m(S)$. For example, suppose the item calls for a real number as the correct response. As the reals are a set of infinite size, the number of elements in S becomes infinite, and $K = m(S) \longrightarrow \infty$. Thus, in the limit, as n increases to infinity,

$$\lim_{n\to\infty} P = \frac{R}{N} - \lim_{n\to\infty} \frac{W/(K-1)}{N} = \frac{R}{N} - \lim_{n\to\infty} \frac{W/(m(S) - 1)}{N}$$

$$= \frac{R}{N} - \lim_{n\to\infty} \frac{W/(n-1)}{N} = \frac{R}{N} - \lim_{n\to\infty} W/(n-1)N$$

$$= R/N \quad \text{(where n is the number of elements in S).}$$

With a large sample, we can estimate the proportion P from data on the highest and lowest 27% of the sample. Let P_{est} be the estimate for P, such that

$P_{est.} = \frac{P_H + P_L}{2}$, where P_H is the proportion of successes in the upper 27%, and P_L in the lower 27% of the sample.

The Davis table transforms the two proportions P_H and P_L into a difficulty index which is on a linear scale; that is, they transform $P_{est.}$ into a standard score and then multiply it by a constant (21.066) and add 50 to this product. This transformation yields an essentially linear scale; a scale of proportions does not constitute such a linear scale. As is true with most data transformations, only difficulty indices based on the same or similar samples are comparable.

By means of the Davis Table (1949) a satisfactory estimate of item difficulty and discrimination which does not require the laborious calculation of the total population and which is easily applied to a series of programed items can be determined.

SUMMARY

This paper presents an analysis of a persistent problem in the development of programed instructional materials: the reduction of data relating to student performance on program frames. It has been customary to use small samples of students and to look at all of their responses in order to make decisions about frame revision. This paper takes the other alternative

as a given, namely, that a large sample of students is used, and asks what approaches to data reduction might be appropriate and useful. The use of a 54% sample to obtain difficulty and discrimination indices is discussed in the light of the problems of programed instruction.

References

Davis, F. G. Item analysis data: Their computation, interpretation, and use in test construction. Harvard Education Papers, No. 2, 1949.

Jacobs, P. I. Some relationships between testing and auto-instructional programing. Audio-Visual communication Rev., 1962, 10, (6), 317-327.

Kelley, T. The selection of upper and lower groups for the validation of test items. J. educ. Psychol., 1939, 30, 17-24.

Wood, Dorothy Adkins. Test construction: Development and interpretation of achievement tests. Columbus, Ohio: Merrill Books, Inc., 1961.